Aalto University

Principles of data protection & ethics in research

Enrico Glerean, Staff Scientist/Data Agent, @eglerean

24/04/2023

Link to the slides + chat https://presemo.aalto.fi/gdpr

15 seconds about Enrico Glerean

Staff scientist at Aalto Scientific Computing

Data agent at Aalto Open Science team

- daily helping researchers with issues related to **computational workflows** with sensitive data, data **minimisation**, preparing legal+ethical forms, HPC… (Research Software Engineer)
- Researchers' trainings on handling personal data in research, research integrity, open science
- research (**neuroimaging**, **medical imaging, experimental psychology**)
- Co-founder of the **Finnish Reproducibility Network**
- Working closely with CodeRefinery (NeIC) and Nordic-RSE

Disclaimer

These topics could cover a full course on their own.

This is just an intuitive introduction to privacy and data protection, data minimisation, concepts from ethics, law, and open science.

Outline

1. Privacy, data protection, and research
2. Ethics
3. Data protection in EU
4. Research compliant with ethics and data protection
5. Reflections on data protection, ethics, and open science

Learning outcomes:

- Understanding why ethics and data protection are important
- Possibilities and limitations of current techniques
- Solutions for ethical and lawful processing of personal data

Let’s start with the references

Where to read and learn more

European Commission “How to complete your ethics self-assessment” (PDF, summary at aalto.fi)

Personal data in research (Aalto)

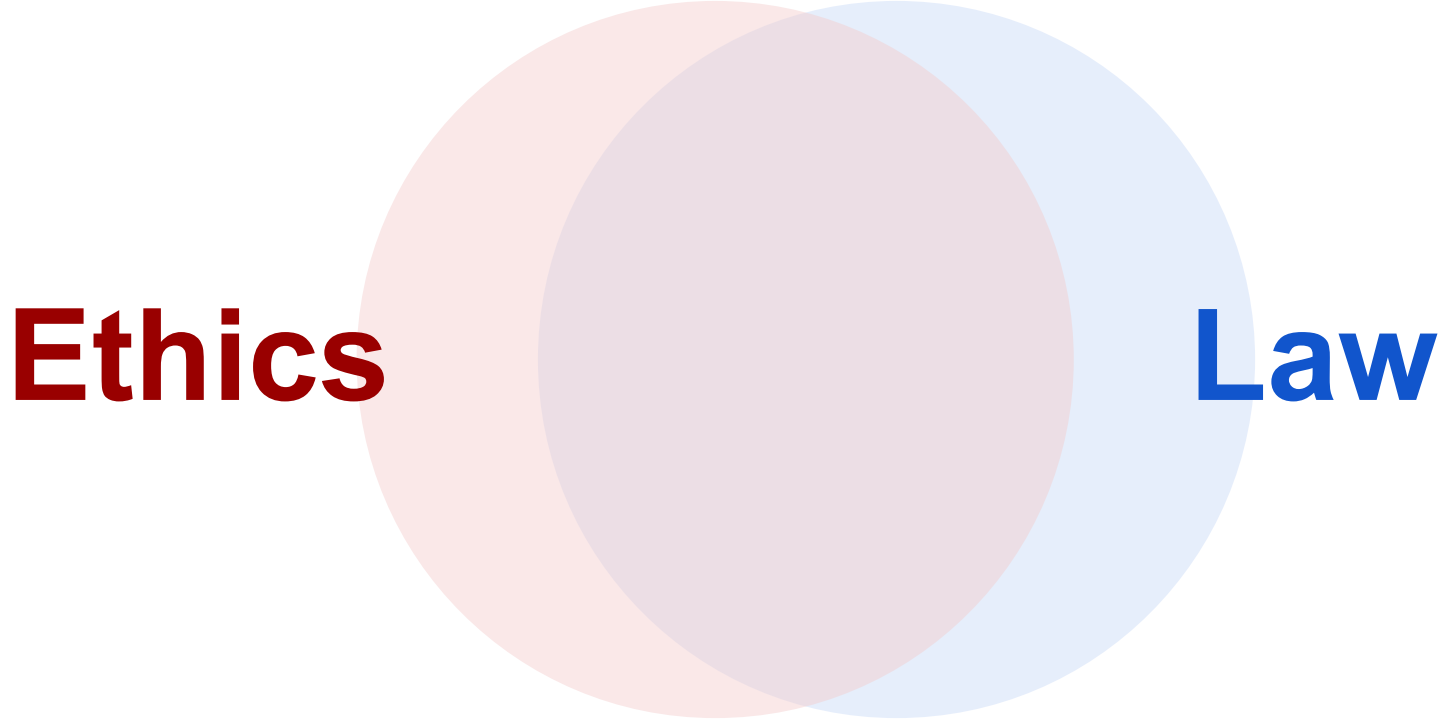
Privacy in Europe (Coursera)

TENK guidelines

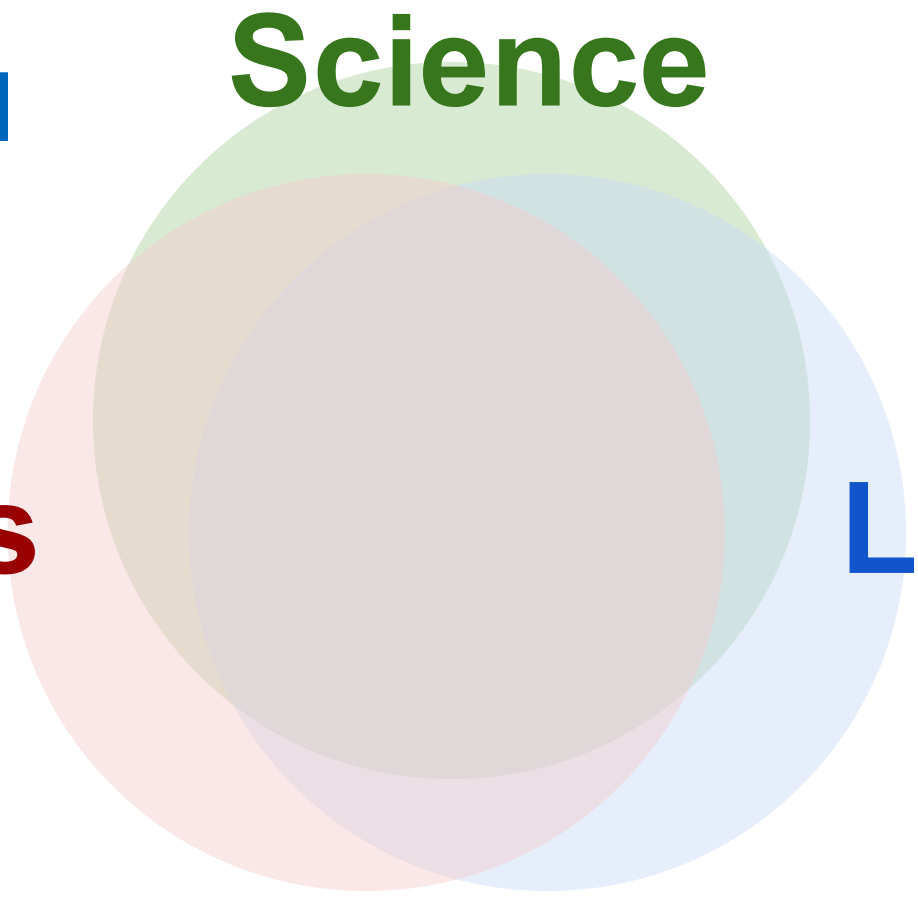
Basics of anonymization

1 Privacy, data protection, and research

Ethics is not Law



Ethics, Law, and Science



Ethics, Law, and Science (in Europe)

Science

The European Code of Conduct for Research Integrity (2011): Principle of Honesty

*"Honesty in developing, undertaking, reviewing, reporting and communicating research in a **transparent**, fair, full and unbiased way"*

Ethics

Universal Declaration of Human Right (1948)

Art. 12 *"No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence,"*

European Convention on Human Rights (1950)

Art. 8 *"You have the right to respect for your private and family life, your home and correspondence."*

Charter of fundamental rights of the european union (2000) Art. 7 *"Everyone has the right to respect for his or her private and family life, home and communications."*

Declaration of Helsinki (1964) Art. 24

"Every precaution must be taken to protect the privacy of research subjects and the confidentiality of their personal information."

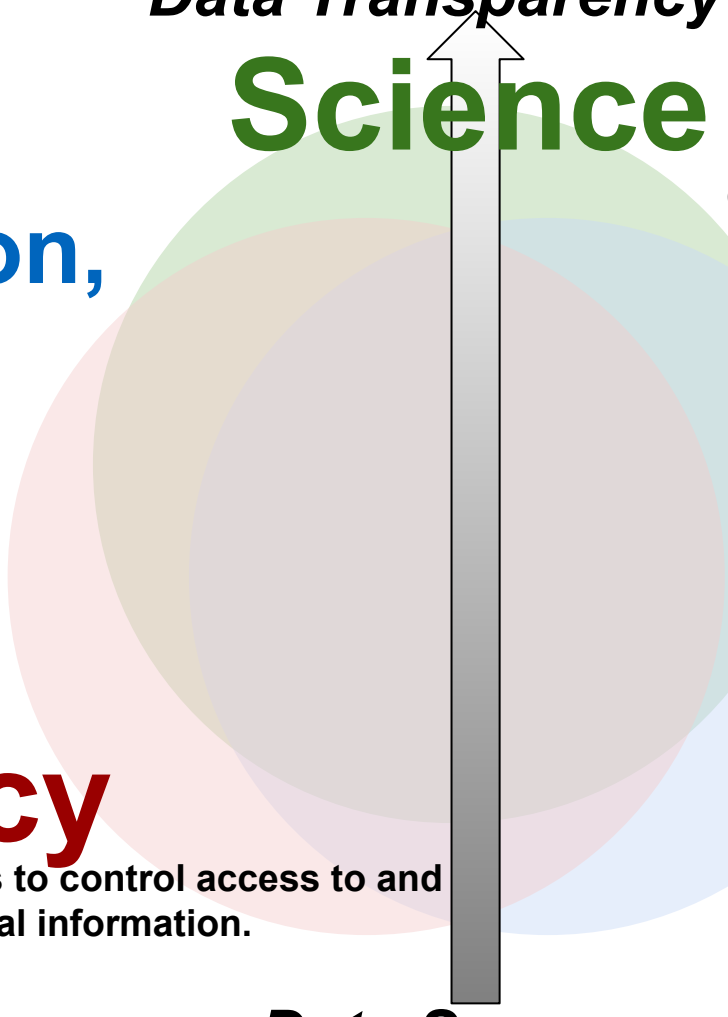
Law

General Data Protection Regulation (2016/679) Art. 5

Lawfulness, fairness and transparency
Purpose limitation
Data minimisation
Accuracy
Storage limitation
Integrity and confidentiality (security)
Accountability

Privacy, Data Protection, Science

Data Transparency



Science

Honesty, transparency, reproducibility, data sharing, data opening

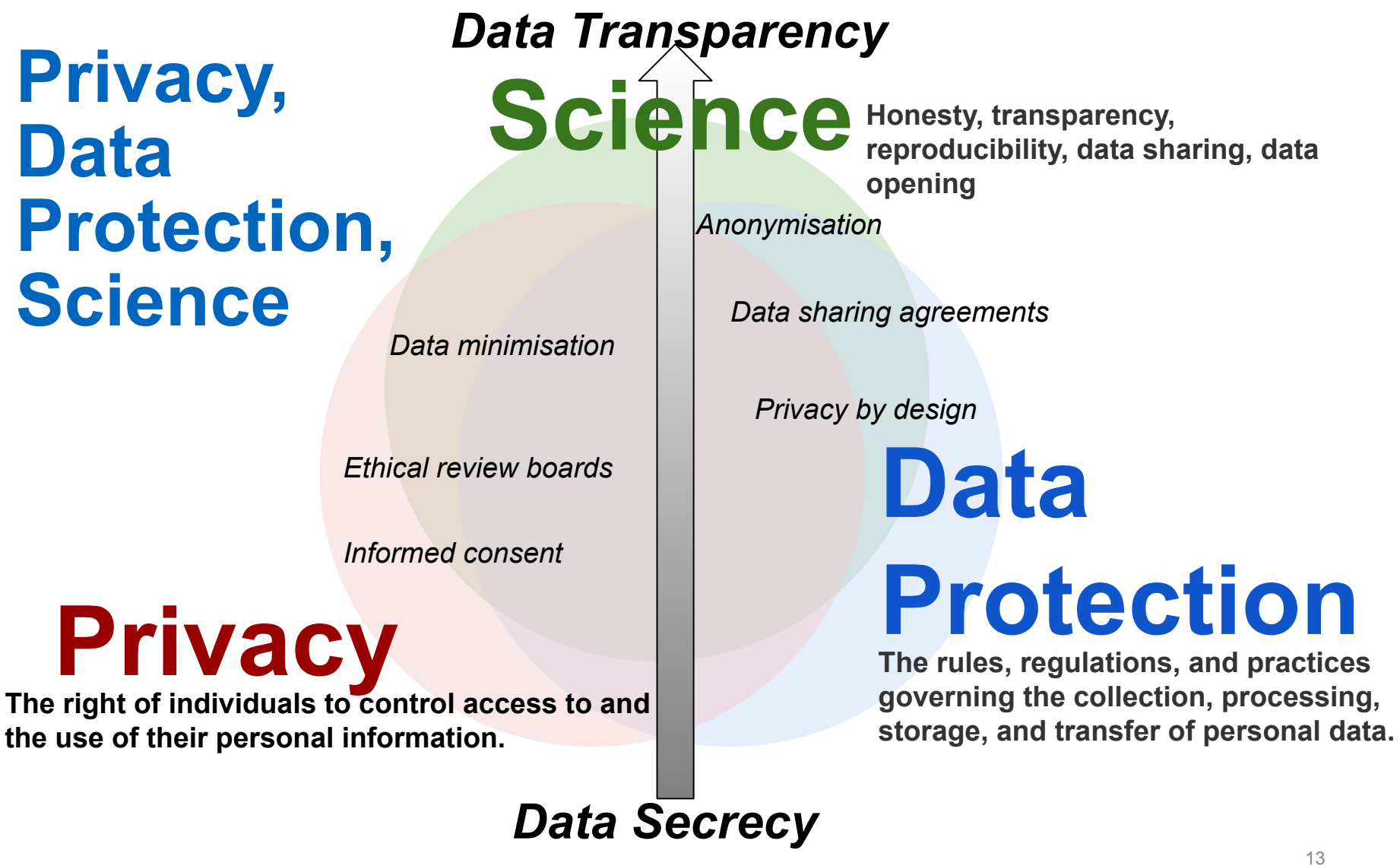
Data Protection

The rules, regulations, and practices governing the collection, processing, storage, and transfer of personal data.

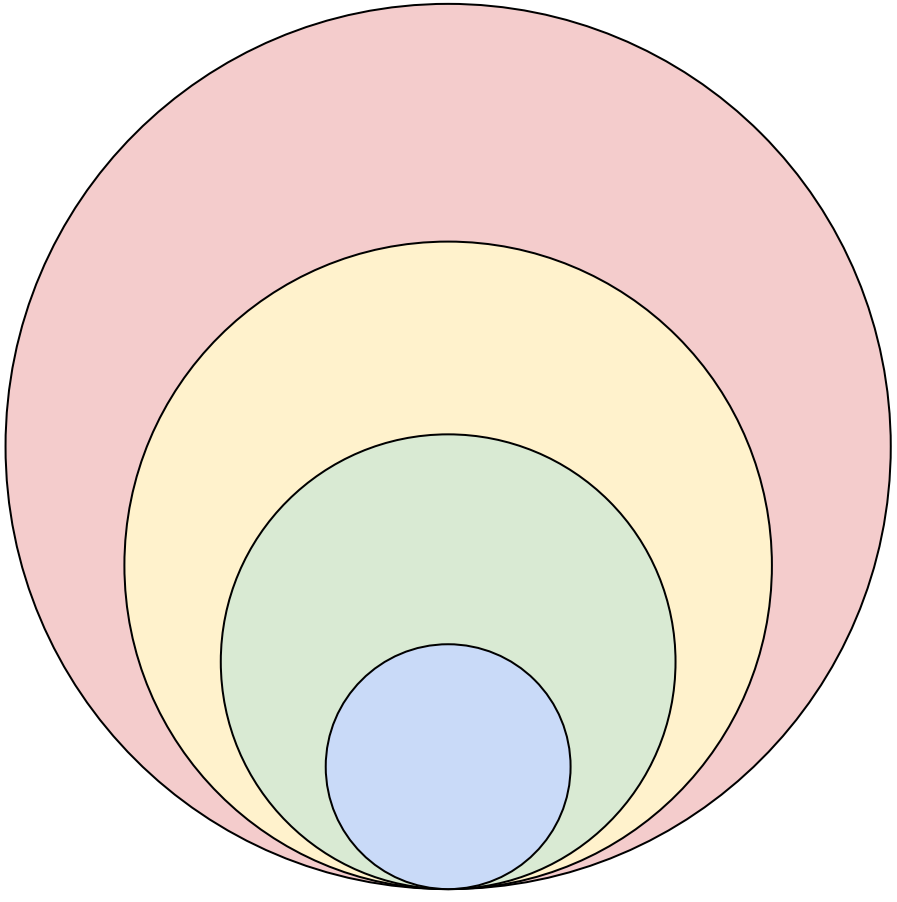
Privacy

The right of individuals to control access to and the use of their personal information.

Data Secrecy


Data Transparency
Privacy, Data Protection, Science
Science
Honesty, transparency, reproducibility, data sharing, data opening
Anonymisation
Data sharing agreements
Data minimisation
Privacy by design
Ethical review boards
Informed consent
Data Protection
The rules, regulations, and practices governing the collection, processing, storage, and transfer of personal data.
Privacy
The right of individuals to control access to and the use of their personal information.
Data Secrecy

2. Ethics



Ethics

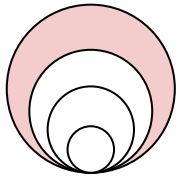
Ethics in Research

Ethics in Research in Europe

Ethics in Research in Finland

Slides available

Ethics



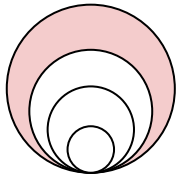
Ethics: norms for conduct that distinguish between **acceptable** and **unacceptable** behavior.

- **Golden Rule:** "*Do unto others as you would have them do unto you*"
- **Hippocratic Oath:** *"First of all, do no harm"*
- **Ten commandments:** *"Thou Shalt not kill…"*

Note: ethics is not just privacy.

https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm

Ethics is not Law



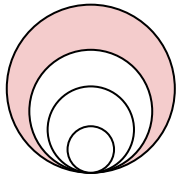
- **Laws enforce accepted moral standards**
- **...but ethical norms are broader and more informal than law**
- **An action may be legal but unethical**
  - Example in pain management: overprescription of opioids
- **An action may be illegal but ethical**
  - Example with helping foreign immigrants stranded at sea
- **This blurred borders are also reflected in ethical considerations for scientific research**

https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm

Ethics: everything can be questioned. With ethical dilemmas **ethics cannot provide an absolute Yes / No answer**.

Law: given a legislation and its agreed interpretation**, law can provide a Yes / No answer.**

Ethics is not Law: the seatbelt analogy



- 1960s: **is it legal** to not wear car seatbelts?
- 1960s: **is it legal** to produce cars without seatbelts?
- 1960s: **is it ethical** to produce cars without seatbelts?
- Today: **is it legal** to not wear car seatbelts?
- Today: **is it legal** to produce cars without seatbelts?
- Today: **is it ethical** to produce cars without seatbelts?
- **Is the law** enforcing us to wear seatbelts **ethical? Is it harming my freedom of choice?**

Law and ethics are changing in time and one can influence another.
Replace seatbelts with “assisted driving”. Or with “wearing masks during a pandemic”. Consider these analogy in the discourse of “freedom of expression” versus “laws against fascism / antisemitism”

Ethics in research

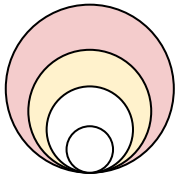
Note: today we are not going to cover research integrity, only ethical issues with study participants / study topics.

Presemo

Think about the studies that you have covered in this course, the studies you might want to run if you will work as a researcher/data analyst in a university or a company.

Is it ethical to …? Is it legal to …?

Ethics in research

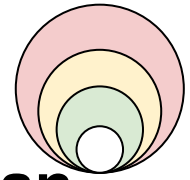


Multiple ethical aspects in research

- collecting, managing, and sharing data
- collegial openness
- ethicality of a study with human subjects
- ethicality of a technology
- misconduct and dealing with misconduct claims

Aalto Research Ethics course (2ECTS) by Henriikka Mustajoki plus other materials at https://mycourses.aalto.fi/course/view.php?id=23138

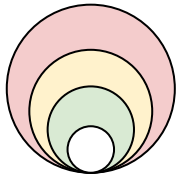
Ethics in research in Europe



1. Medical research with human participants, human organs, tissues and cells

‘Medical research means research involving intervention in the integrity of a person, human embryo or human foetus for the purpose of increasing knowledge of health, the causes, symptoms, diagnosis, treatment and prevention of diseases or the nature of diseases in general, and which is not a clinical trial as defined in the Clinical Trials Regulation’ (Medical Research Act (488/1999), *Section 2(1) as amended by the Act 984/2021, see also below).*

Ethics in research in Europe

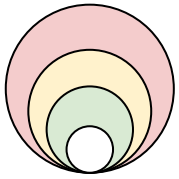


2. Non-medical research with human participants

- Think of all the risks which may occur to your research participants and how to mitigate them. **The risks could be physical, social and psychological**.
- Familiarise yourself and comply with personal data regulations.
- Aalto University has a **Research Ethics Committee** which handles Aalto researchers' requests for research ethics reviews for research projects with human participants.

This is **not** an ethical permit, this is a statement of ethicality from the committee. While sometimes it is not mandatory to go through ethical review in Finland, you might encounter difficulties with the journals when publishing your findings.

Ethics in research in Europe

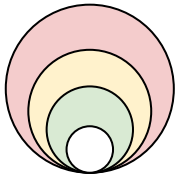


3. Processing of personal data

- Personal data is any data or information relating to an identified or identifiable natural person.
- Basically anything you measure from an individual is a fingerprint.
- True-anonymisation is very hard and often impossible without completely destroying the data.
- Ethical review needed if **special categories of personal data** are processed: *Data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orient at ion (Article 9(1) of the GDPR)*

More on this later in the slides

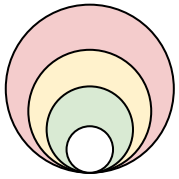
Ethics in research in Europe



4. Research with animals

- You need a permit to work with test animals from your local Regional State Administrative Agency. The space/laboratory/project needs to have a licence to acquire, breed, keep and take care of animals for scientific or educational purposes.
- If you work with genetically manipulated test animals, you need a permit from the Board for Gene Technology.
- A research ethics review is often required when doing research with animals.

Ethics in research in Europe

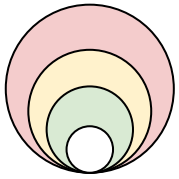


5. Research in non-EU countries, global south/developing countries

- Humans and vulnerability: communities, subjects and vulnerabilities, agency distribution, power dynamics
- Materials of historical value: architecture, historical landmarks, cultural sensitivity, intangible cultural heritage
- Sensitive topics: genetic, health, sexual, lifestyle, ethnicity, political opinion, religious or philosophical conviction, value conflicts
- Personal data transfers

Note that you might need a research ethics review

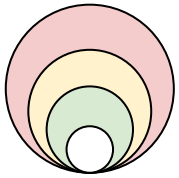
Ethics in research in Europe



6. Environment & safety

- The European Commission also requires that you follow the **'do no significant harm'** principle which means that your research should not do any significant harm against any of the six environmental objectives covered by the Taxonomy Regulation (2020/852):
  a. Significant greenhouse gas emissions
  b. Significantly harming climate change
  c. Significant harm to water and marine resources
  d. Significant harm to recycling
  e. Significant increase in pollution
  f. Significant harm to protection of biodiversity

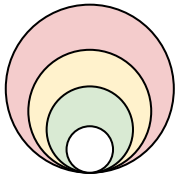
Ethics in research in Europe



7. Artificial Intelligence

- Human agency and oversight
- Privacy and data governance
- Fairness, diversity and non-discrimination
- Accountability
- Transparency
- Societal and environmental well-being (impact)

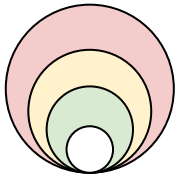
Ethics in research in Europe



8. Dual use items

- dual-use items: non-military items that can be modified to be used in e.g. terrorism or as weapons of mass destruction
- technology which can be used to undermine democracy or human rights
- other items subject to regulations, you need to apply the proper procedures and apply for permits

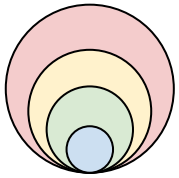
Ethics in research in Europe



9. Potential misuse of results

- Misuse looks at the potential **unethical use of your research in broad perspective**. What can you do to avoid any of your research materials (e.g. biological, chemical, radiological and nuclear security-sensitive materials and explosives) or **outcomes from being used for unethical purposes**, ending up in the wrong hands, or for example being used to violate human rights or to undermine democratic processes

Ethics in research in Finland



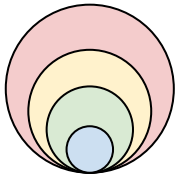
Finnish National Board on Research Integrity (TENK)

Three areas of ethical principles:

1. **Respecting the autonomy of research subjects**
2. **Avoiding harm**
3. **Privacy and data protection.**

https://www.tenk.fi/en/ethical-review-in-human-sciences

Ethics in research in Finland



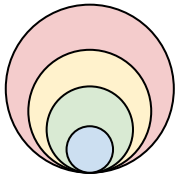
Finnish National Board on Research Integrity (TENK)

Three areas of ethical principles:

1. **Respecting the autonomy of research subjects**
   a. Voluntary participation
   b. Autonomy and research involving minors
   c. Autonomy and age limits
   d. Information for subjects
   e. Exceptions from informed consent (needs EthR)

https://www.tenk.fi/en/ethical-review-in-human-sciences

Ethics in research in Finland



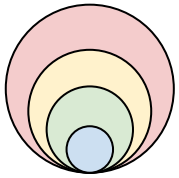
Finnish National Board on Research Integrity (TENK)

Three areas of ethical principles:

1. **Respecting the autonomy of research subjects**
2. **Avoiding harm**
   a. Avoiding mental harm
   b. Avoiding financial and social harm
   c. Studies containing risks of harm (needs EthR)
3. **Privacy and data protection.**

https://www.tenk.fi/en/ethical-review-in-human-sciences

Ethics in research in Finland



Finnish National Board on Research Integrity (TENK)

Three areas of ethical principles:

1. **Respecting the autonomy of research subjects**
2. **Avoiding harm**
3. **Privacy and data protection**
    a. Protecting research data and confidentiality
    b. Storing or destroying research data
    c. Protecting privacy in research publications

https://www.tenk.fi/en/ethical-review-in-human-sciences

When should we apply for ethical review? (non medical studies with human participants)

TENK Guidelines Section 4.2

https://www.tenk.fi/sites/tenk.fi/files/Ihmistieteiden_eettisen_ennakkoarvioinnin_ohje_2019.pdf

The researcher must request an ethical review statement from a human sciences ethics committee, if their research contains any of the following:

a) Participation in the research **deviates from the principle of informed consent**,

b) the research involves **intervening in the physical integrity** of research participants,

c) the focus of the research is on **minors under the age of 15**, without separate consent from a parent or carer or without informing a parent or carer in a way that would enable them to prevent the child's participation in the research,

d) research that exposes participants to **exceptionally strong stimuli,**

e) research that involves a risk of **causing mental harm** that exceeds the limits of normal daily life to the research participants or their family members or others closest to them or

f) conducting the research could involve a **threat to the safety of participants** or researchers or their family members or others closest to them.

3 Data protection in the European Union

What is the GDPR?

Established in 2018, **the General Data Protection Regulation (GDPR)** is a **groundbreaking data protection regulation** that revolutionized the way **personal data** is handled in the EU, granting individuals greater **control** over their information and holding organizations **accountable** for responsible data processing.

https://tietosuoja.fi/en/home

What is personal data?

- **Personal data is a broad concept under the EU's General Data Protection Regulation**
- **"Personal data" is <u>any data about living people from which they can be identified</u>**
  - If you collect <u>information **from** or **of** persons</u>, consider it as **personal data**
  - Exception: **anonymous data**

Personal data: Direct and indirect identifiers

a) **Direct identifiers:** information which is **sufficient on its own to identify an individual**

- e.g. a person's name, email address (containing the person's name), personal identification number, fingerprints, a facial image, a person's voice, video, brain scan images, dental records, DNA

b) **Strong indirect identifiers:** information which **can be used to identify an individual fairly easily**

- e.g. a postal address, a phone number, a vehicle registration number, bibliographic citation of a publication, an email address not in the form of the personal name, an unusual job title, a very rare disease, a job position held by only one person at a time, a student ID number, a bank account number, IP address of a computer, cookie identifier, RFID tags, location data

c) **Indirect identifiers:** information that on its own is not enough to identify someone but, when **linked with other available information, could be used to deduce the identity of a person:**

- e.g. age, gender, education, status in employment, economic activity and occupational status, socio-economic status, household composition, income, marital status, mother tongue, ethnic background, place of work or study, postal code, municipality, major region.

Special categories of personal data (art 9)

- personal data revealing racial or ethnic origin;
- personal data revealing political opinions;
- personal data revealing religious or philosophical beliefs;
- personal data revealing trade union membership;
- genetic data, biometric data (where used for identification purposes);
- data concerning health;
- data concerning a person's sex life and data concerning a person's sexual orientation.

The principles of the GDPR (art 5)

GDPR Principle	Practical Actions in Research Projects
1. Lawfulness, fairness, and transparency	Obtain and manage informed consent from research participants Clearly explain data processing activities, purposes, and potential risks to participants Maintain transparency in data sharing, storage, and access policies
2. Purpose limitation	Specify the research objectives and data processing purposes before collecting personal data Use collected data only for the stated purposes Obtain additional consent or establish a new legal basis if the purpose changes
3. Data minimization	Collect only the minimum amount of personal data necessary to achieve research objectives Employ data anonymization or pseudonymization techniques to reduce the scope of identifiable information
4. Accuracy	Implement processes to ensure the accuracy and up-to-date nature of personal data Allow research participants to rectify inaccurate or incomplete information Regularly review and update collected data as needed
5. Storage limitation	Establish retention periods for personal data based on research objectives and legal requirements Regularly delete or anonymize personal data when it is no longer necessary for the stated purpose or when the retention period ends
6. Integrity and confidentiality	Use appropriate security measures to protect personal data from unauthorized access, disclosure, or misuse Train research team members on data protection and privacy practices Implement access controls and encryption for data storage and transfer
7. Accountability	Document data processing activities, legal bases, and compliance measures Appoint a data protection officer (DPO) when required Conduct data protection impact assessments (DPIAs) for high-risk processing activities Demonstrate adherence to GDPR principles and requirements in research practices

The legal bases of the GDPR (art 6)

Legal Basis	Applicability in Research and Other Contexts
1. Consent	Obtaining explicit and informed consent from research participants for data collection and processing Marketing activities *Consent can be withdrawn at any time; Reusing data for new purposes may require obtaining new consent*
2. Contract	Processing personal data in the context of a contract (e.g., employment, service provision) *Limited to the specific terms and purposes outlined in the contract*
3. Legal Obligation	Processing personal data to comply with legal requirements (e.g., tax reporting, anti-money laundering regulations) *Must be based on a specific legal obligation under EU or national law*
4. Vital Interests	Processing personal data to protect the life or health of an individual (e.g., emergency medical situations) *Rarely applicable in research settings*
5. Public Task	Public authorities processing personal data for official tasks or public interest Research conducted for public interest *Must be based on EU or national law. Greater flexibility for data reuse in scientific or archival purposes*
6. Legitimate Interests	Processing personal data when necessary for the legitimate interests of the organization or a third party (e.g., fraud prevention) *Must be balanced against the rights and interests of data subjects. Rarely applicable in research settings*

Rights of the data subject (art 12-23)

According to the General Data Protection Regulation (GDPR), data subjects have the right

- **to obtain information on the processing of their personal data**
- **of access to their data**
- **to rectification of their data**
- **to the erasure of their data and to be forgotten**
- **to restrict the processing of their data**
- **to data portability**
- **to object to the processing of their data**
- **not to be subject to a decision based solely on automated processing.**

4 Reconcile data protection and transparency in research

Aalto University
School of Science

Measures to ensure data protection and ethicality in research

- **The Ethical review process**
- **The ethical consent to participate**
- **Informing the data subjects (privacy notice)**
- **Data minimisation** (pseudonymisation, anonymisation)
- **Privacy by default**
- **Secure data processing strategies** when minimisation is not possible or risks are high (hundreds of subjects, minors (DPIA)) or when required by legislation (secondary use of health data)

Data minimisation: pseudonymization, anonymization

Minimisation:

- Only the **minimum amount of personal data necessary to accomplish a task** (e.g. research) should be collected. Personal data must not be collected just in case they might be useful in the future. There has to be a **clear, specified need for collecting the personal data**.

Note that **there are no restrictions** on what type of personal data you can collect, however extra care must be taken when collecting **special categories of personal data GDPR Art. 9**

(personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation)

Data minimisation: pseudonymization, anonymization

Pseudonymisation :

- **Replacement** of identifiers with pseudonyms or codes, which are kept separately and protected by technical and organisational measures
- **the data are pseudonymous (and hence personal data) as long as the additional identifying information exists**

“We make it difficult for others to re-identify, but we still have the key”

Example: registry data, medical data, longitudinal studies

Data minimisation: pseudonymization, anonymization

~~De-identification~~ (not a good term to use!):

- **removal of direct identifiers** (Elliot et al. 2016)
- **the data allow re-identification with additional data (and hence they are personal data)**

"We make it difficult for others to re-identify, we even throw the key, but there is a way to re-identify"

E.g. a picture of a fingerprint, a "defaced" brain scan

Data minimisation: pseudonymization, anonymization

Anonymisation:

- to anonymise personal data means to **irreversibly remove identifying information** from the data so that **a person cannot be identified based on the data**
- all the means **"reasonably likely"** to be used for the identification of individuals must be considered when assessing whether the data has been anonymised (also **information available from other data sources shall also be taken into account)**
- **The GDPR does not apply for truly anonymous data (GDPR recital 26)** … but what about ethics?

Data minimisation spectrum

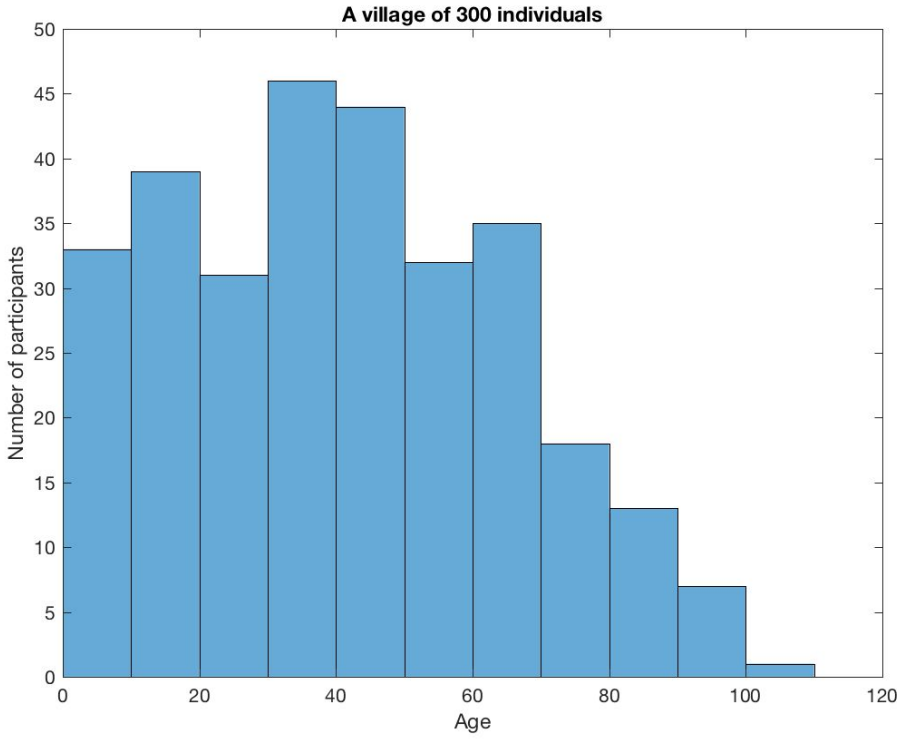
Personal data with direct identifiers and no minimisation	**Pseudonymisation** Replacing strong identifiers: tokenization, hashing, encryption	**Masking** Suppression of strong identifiers	**Anonymisation** K-anonymity, l-diversity, t-closeness, perturbation (data swapping, differential privacy)	**Anonymisation** Aggregation, data synthesis

Easy to re-identify	**Can be re-identified with the key**	**The key is lost, but other data (will) exist**	**Impossible to re-identify, but still related to an individual**	**Impossible to re-identify and not related to an individual anymore**

4.2 Understanding anonymisation through re-identification

Aalto University
School of Science

Singling out – a small village with 300 individuals

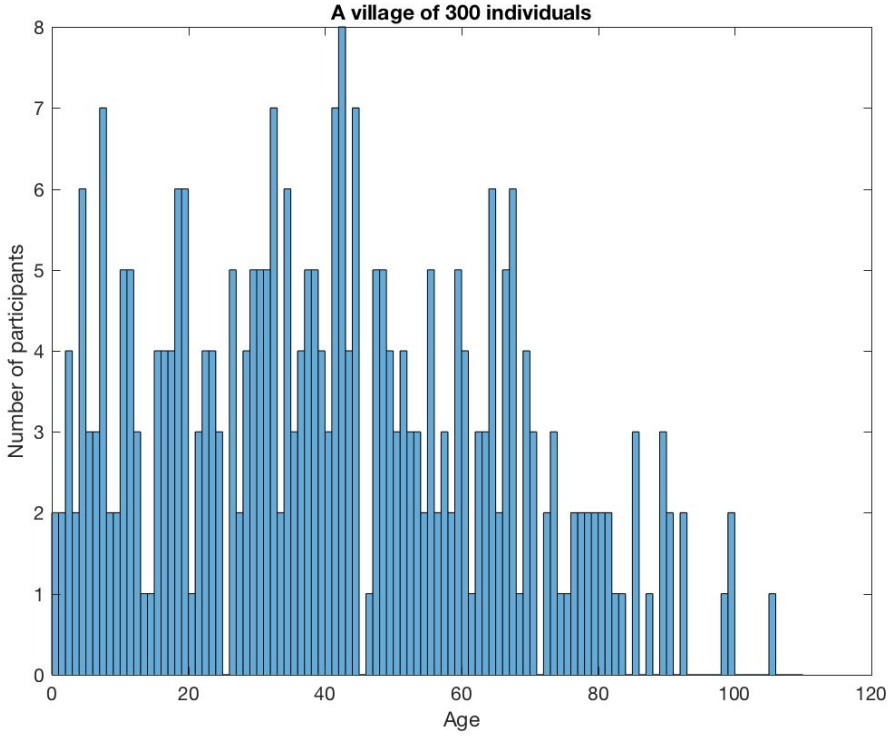


- People counted together based on their “decade” age
- We can single out one very old individual which most likely everybody in the village knows

ID	AGE group	Sensitive info (e.g. “was in jail”)
1	50	0
2	40	0
3	**100**	1
...	40	1

Red = only one in the dataset

Singling out – a small village with 300 individuals

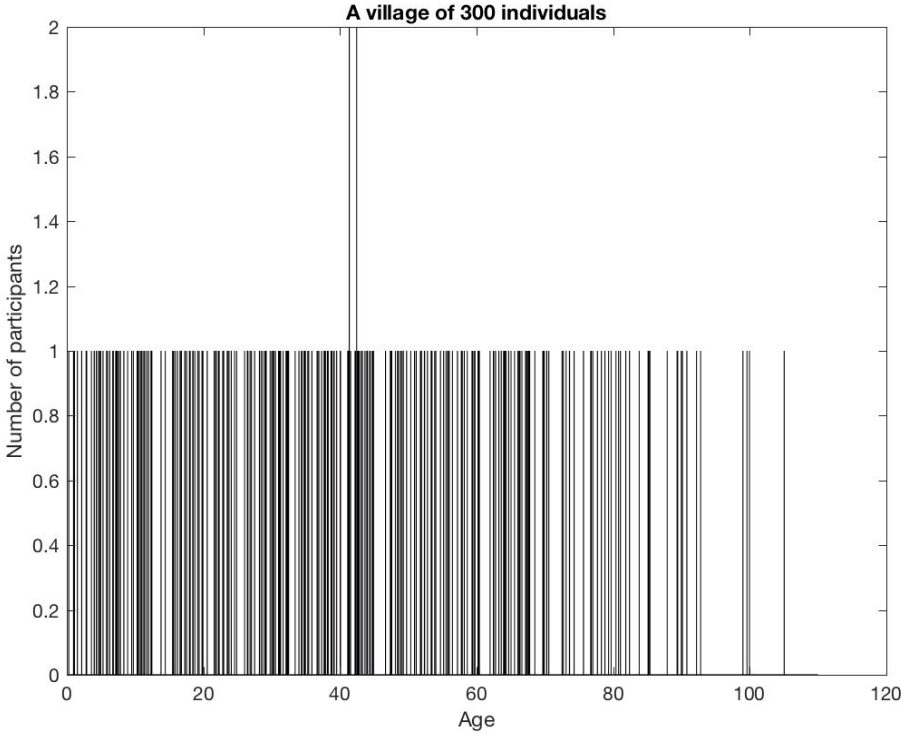


- People counted together based on their age as integer year
- Increasing granularity of data, makes the subjects more identifiable.

ID	AGE group	Sensitive info (e.g. “was in jail”)
1	56	0
2	43	0
3	**106**	1
...	**46**	1

Red = only one in the dataset

Singling out – a small village with 300 individuals

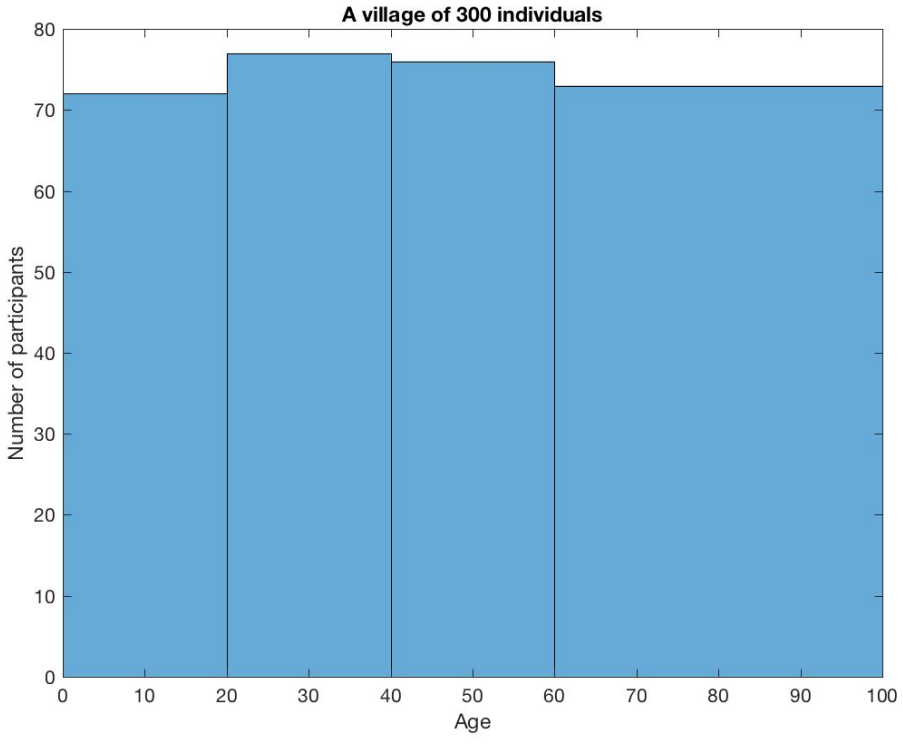


- People counted together based on their age including day and time of birth
- We can single out basically every individual

Increasing granularity of data, makes the subjects more identifiable.

K-anonymity = 1

Solution: binning the data into uniform sub-groups



- Groups that are less represented should be merged together

This anonymisation method is called **generalisation**

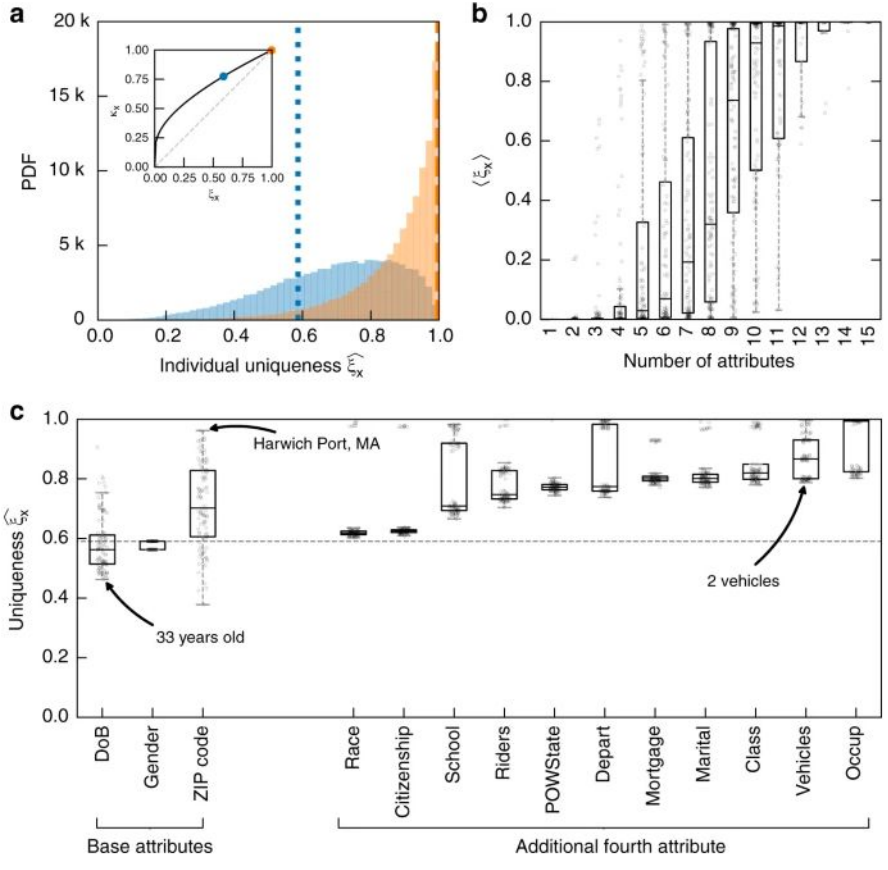
K-anonymity =~ 70

Each person contained in the release cannot be distinguished from at least k - 1 individuals whose information also appear in the release.

Singling out – more data, lower k-anonymity

- With ~4 pieces of information you can uniquely identify a character from “guess who?” board game
- Even though each piece of information could be generalised into a broad category, more information makes the individual more identifiable.

Singling out – more data, lower k-anonymity



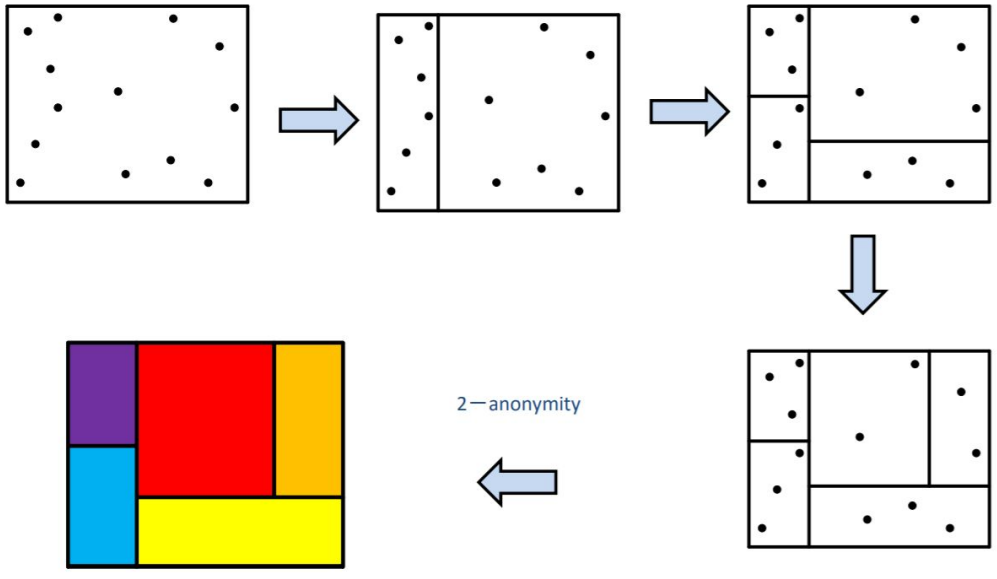
- With ~15 pieces of information you can uniquely identify 99.98% of the American population.

"Estimating the success of re-identifications in incomplete datasets using generative models" Rocher et al (2019)
https://www.nature.com/articles/s41467-019-10933-3

Re-identification from fingerprinting

- **When you start collecting more rich data from individuals** (basically anything related to their body: fingerprints, DNA, eye movements, brain activity, brain morphology, electrocardiogram, gait, voice, face traits) **anonymisation becomes impossible, unless you want to make the data unusable.**
- **Data minimisation is still necessary** (e.g. removing direct identifiers), and data need to be handle as personal data

K-anonymity to more identifiers



Mondrian method (LeFevre et al 2006). Figure from these slides

- When more than one identifier is considered, one need to find **subgroups**.
- The more identifiers are added the more it is impossible to achieve k-anonymity without losing data
- Same issues as other clustering or dimensionality reduction methods. If the clusters / principal components do not make much sense, it is difficult to claim something generizable about the findings.

Re-identification attacks and other risks

- **Privacy attacks** (https://dl.acm.org/doi/abs/10.1145/3436755)
  - Inference attacks (by analysing the data one can infer information about an individual)
  - Linkage attacks (e.g. the famous linkage between two public datasets from Netflix and IMDB)
  - Model extraction attacks (various types, e.g. the original face extracted from a ML model)
- **When studying a rare population, there are higher chances for the participants to be re-identified**

  Individuals with rare diseases, individuals that are famous (musicians, celebrities, politicians)

4.3 New challenges posed by AI

Example of model extraction attack from 2023 using text to image models (stable diffusion)

- https://arxiv.org/pdf/2301.13188.pdf

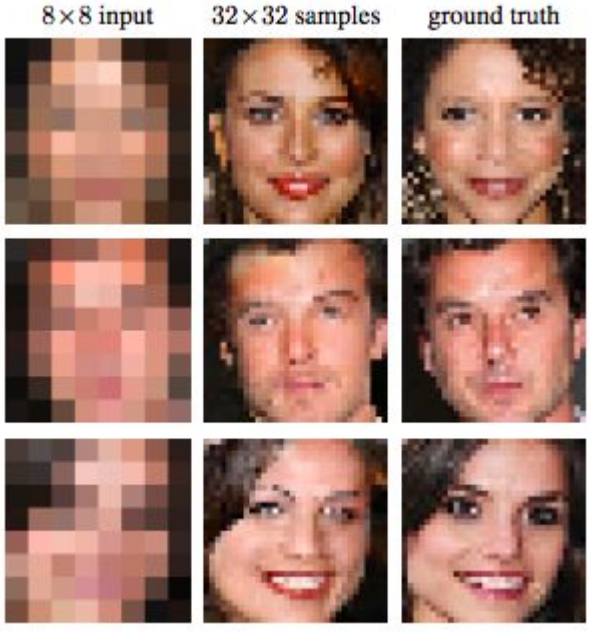
Figure 3: Examples of the images that we extract from Stable Diffusion v1.4 using random sampling and our membership inference procedure. The top row shows the original images and the bottom row shows our extracted images.

Re-identification through synthetic data

Advances in machine learning technology can potentially allow re-identification of minimised data.

- **Reconstructing faces from low quality images:**

  https://arxiv.org/pdf/1702.00783.pdf



Re-identification through synthetic data

Advances in machine learning technology can potentially allow re-identification of minimised data.

- **Reconstructing faces from defaced MRIs of the head.**

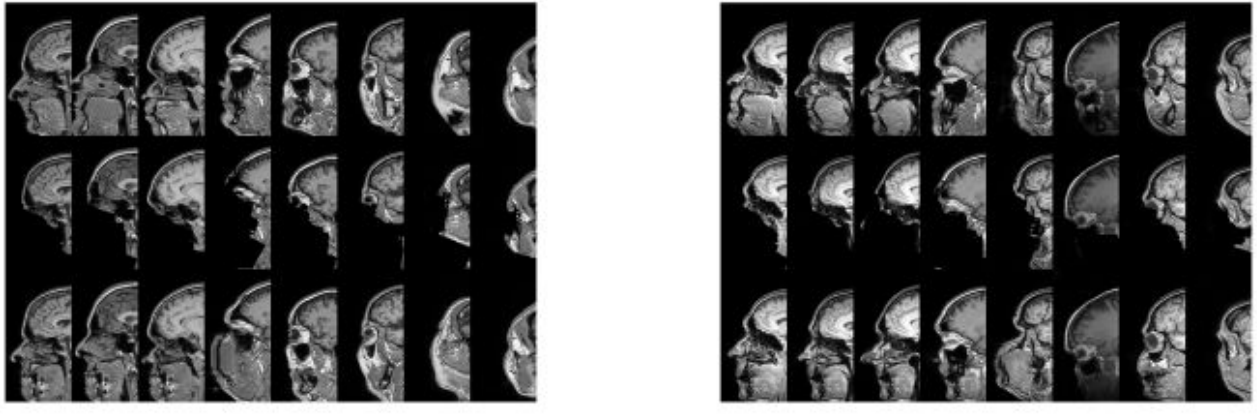


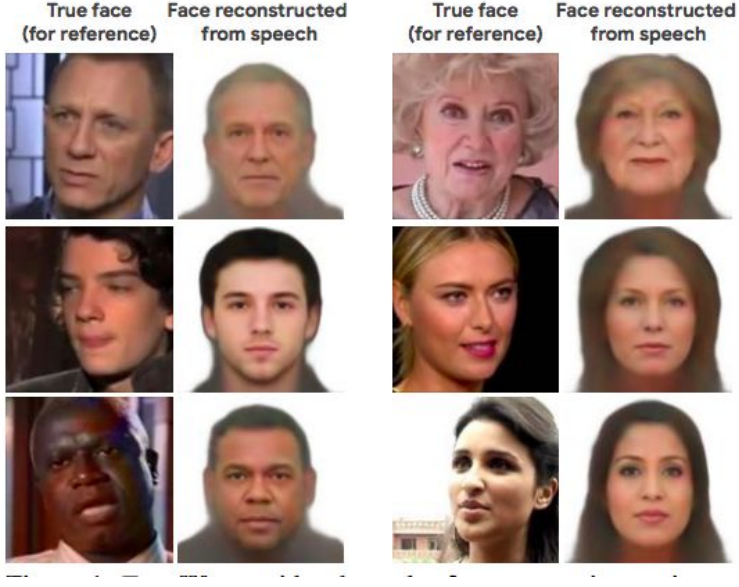
Fig. 2. Typical results of refacing face-removed images. Left: results for training using only subjects from Guy's hospital, Right: results for training using data from all 3 sites. Top row: original image, middle row: face-removed image, bottom row: reconstructed image. CycleGAN learns to add a face, but in many cases it is not the correct face.

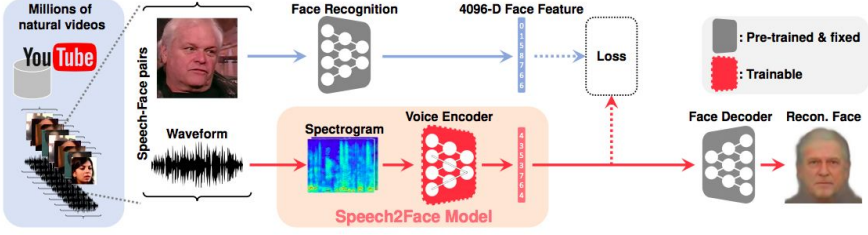
http://www.diva-portal.org/smash/get/diva2:1359959/FULLTEXT01.pdf

Re-identification through synthetic data

Advances in machine learning technology can potentially allow re-identification of minimised data.

- **Reconstructing faces from voice recordings**





https://openaccess.thecvf.com/content_CVPR_2019/papers/Oh_Speech2Face_Learning_the_Face_Behind_a_Voice_CVPR_2019_paper.pdf

4.4 Anonymisation in practice

Anonymisation before data collection: Privacy by design

- Plan your research with **minimisation** in mind
  - What is the minimum amount of personal data that I need to answer my research question?
  - Justify your data/protocol whether you have an hypothesis or explicitly mention that it is exploratory (preregistration and DMP)
  - Structured data in favour of unstructured data (i.e. avoid open ended questions)
- **Data Management Plan** should mention about your data minimisation strategy.
- **However, consider also the ethical aspects:** e.g. you might collect health data and come across an *incidental finding*, you still need to be able to contact the participant before minimising the data

Anonymisation before data collection: Privacy by design with background data

- **Restrict as much as possible the range of answers**
- **Do not promise that data will be fully anonymous** if you do not have proof to show the participant that it will be truly anonymous or if you do not have an anonymisation strategy to show.

In doubt, just tell them that you are collecting personal data and provide a privacy notice along with the consent to participate to your study.

Anonymisation after data collection

- **Participants' background information**: Destroy, obfuscate, generalise background variables according to the table at https://www.fsd.tuni.fi/en/services/data-management-guidelines/anonymisation-and-identifiers/
- **Health data**: minimise file headers, remove direct identifiers (e.g. faces from MRI)
- **Geospatial data:** https://www.sciencedirect.com/science/article/pii/S0198971520302465#f0010
- **Audio visual data**
- **…**
- **What is your data?**

Identifier type	Direct identifier	Strong indirect identifier	Indirect identifier	Anonymisation method
Personal identification number	x			Remove
Full name	x			Remove/Change
Email address	x	x		Remove
Phone number		x		Remove
Postal code			x	Remove/Categorise
District/part of town			x	Categorise
Municipality of residence			x	Categorise
Region			x	(Categorise)
Major region			x	
Municipality type			x	
Audio file	x			Remove
Video file displaying person(s)	x			Remove
Photograph of person(s)	x			Remove
Year of birth		x		Categorise
Age			x	Categorise
Gender			x	
Marital status				

https://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html

Anonymisation after data collection: some tools

- https://amnesia.openaire.eu/ (tabular data)
- https://arx.deidentifier.org/ (tabular data)
- https://sourceforge.net/projects/anony-toolkit/ (tabular data)
- https://surfer.nmr.mgh.harvard.edu/fswiki/mri_deface (MRI)
- http://mist-deid.sourceforge.net/ (unstructured medical records)
- https://nlp.stanford.edu/software/CRF-NER.html (NLP for unstructured text)

Anonymisation in qualitative research

- **Various qualitative methods produce data in the form of text** (interview transcripts, field notes)
- **The data can contain all sorts of direct and indirect identifiers**
- **Manual minimisation (anonymisation) follows these rules:**
  a. Replacing personal names with aliases
  b. Categorising proper nouns
  c. Changing or removing sensitive information
  d. Categorising background information
  e. Changing values of identifiers
- **Advances in ML allow for NER (Name Entity Recognition).**
  See a recent example https://arxiv.org/abs/2208.13081 (1% de-anonymisation rate)

https://www.fsd.tuni.fi/en/services/data-management-guidelines/anonymisation-and-identifiers/#anonymising-qualitative-data

Name Entity Recognition

- Subtask of natural language processing (**NLP**) that involves identifying and classifying named entities in unstructured text. (ref)
- Named entities are typically **proper nouns**, such as names of people, **organizations**, **locations**, **dates**, and other specific items e.g. **special categories of personal data**.

  Personal data revealing racial or ethnic origin; Political opinions; Religious or philosophical beliefs; Trade union membership; Genetic data and biometric data processed for the purpose of uniquely identifying a natural person; Data concerning health; Data concerning a natural person's sex life or sexual orientation.

- The goal of NER is to **automatically detect these entities** and categorize them into predefined classes.
- **Different implementations according to fields** (a medical NER system might focus on medical terms, a legal NER on legal cases etc)
- Practical tips: it should **run locally**, it can be **challenging to set up**, it will always require **manual quality control**

Anonymisation in qualitative research: practical tips

- **Good planning leads to excellent anonymisation**
- **Do not anonymise like a robot, decide before starting what is relevant**
  a. E.g. family and social relationships are important for the research question, then terms like "mother" "sister" "neighbour" should stay and when the names of these individuals appear in full, they can be replaced with the appellative.
  b. Counter-example: family relationships do not matter, so "mother" and "sister" could both be replaced with "a close relative".
  c. Delete everything that is not relevant
- **It is challenging to prove that the anonymised text is truly anonymous, especially when linkage attacks are possible (e.g. anonymised tweets can be re-identify)**
- **When possible, use coding of the interviews + concepts, but remember that you might still end up with fingerprints**

4.5. Workflows when anonymisation is not possible

When anonymisation is not possible

- **Adopt secure workflows** (use remote computing, encryption)
- **Keep the data in the safest place** (e.g. SECDATA, CSC SD)
- **Bring the code to the data**
  a. Code and data on same system
  b. Federated analysis approach (usually via containers)
- **Bring (part of) the data to your code**
  a. Subpart of the data e.g. Beacons in genomics (https://www.nature.com/articles/s41587-019-0046-x)
  b. Synthetic data with same statistical properties of the data you work with (https://arxiv.org/abs/1912.04439)

Federated analysis approaches

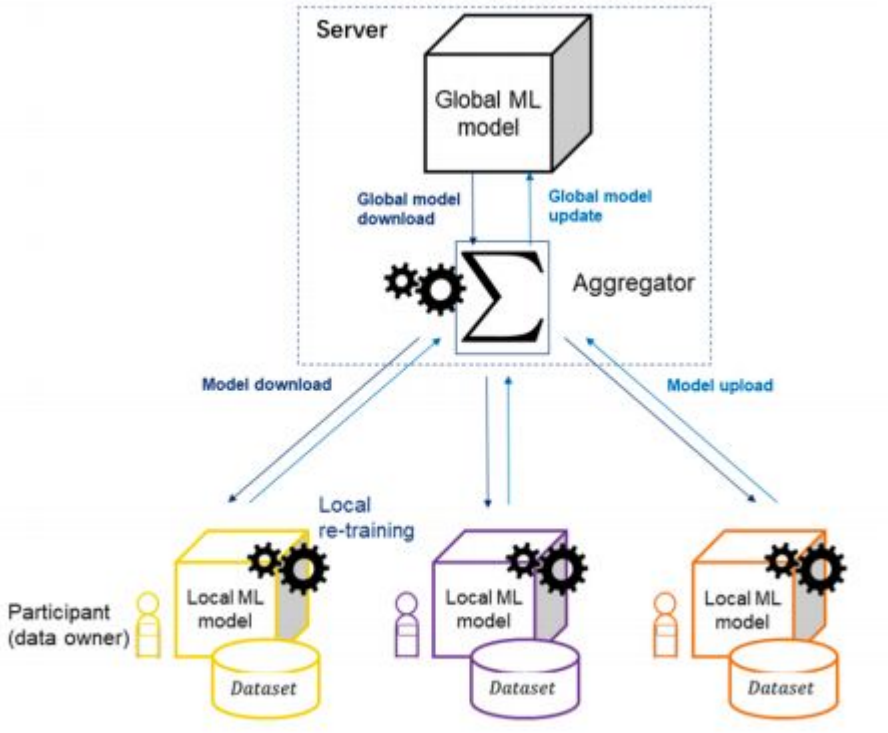
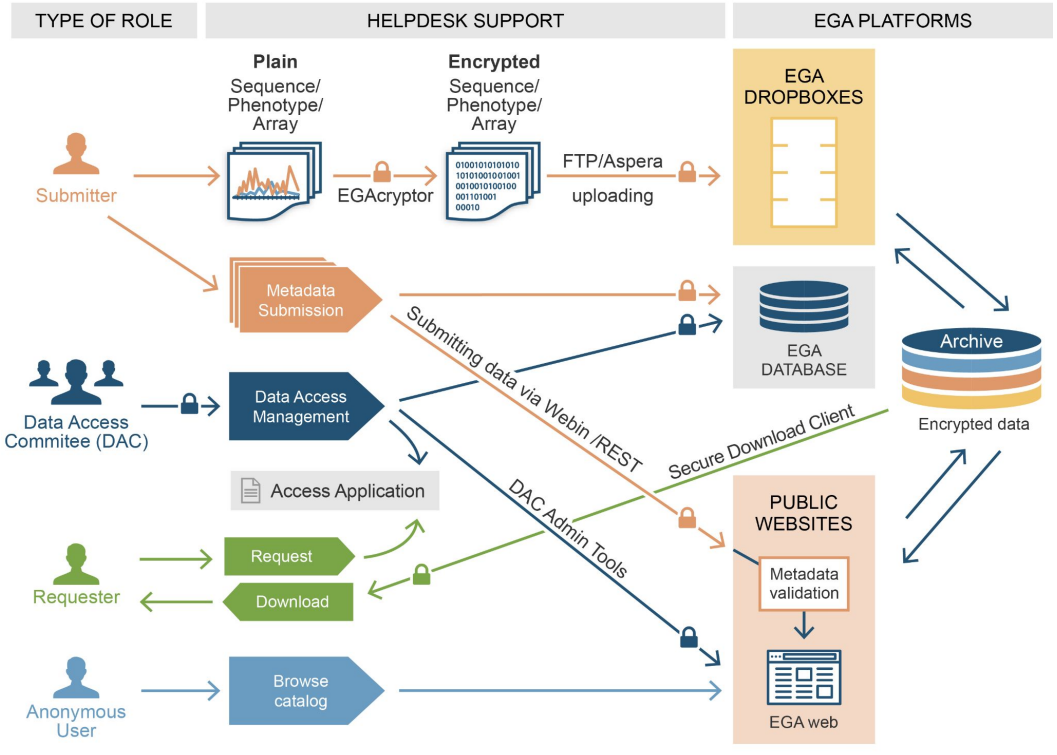


Figure 4: Federated Learning Architecture (client-server FL)

- Data stays with owners who can run the same code
- Aggregator can join models from multiple data owners

Figure from https://aaltodoc.aalto.fi/bitstream/handle/123456789/43561/master_Xia_Yuxi_2020.pdf?sequence=1&isAllowed=y

Data access control



Procedure adopted by the (federated) EGA (European Genome Phenome Archive)

In the federated approach, "Download" is replaced with remote secure computing

Figure from https://ega-archive.org/about/introduction

Other solutions

- **Confidential computing (encrypted containers on secure clusters)**
- **Differential privacy (“perturb” the data while keeping their covariance structure)**
- **Data synthesis (work on a simulated dataset from the original data)**
  - https://github.com/DPBayes/twinify

5. Reflections on anonymisation, ethics, and open science

Data minimisation: start with your own data

Understand your own privacy, your personal data, evaluate the risks that you expose yourself to. Europe is the best place to be!

- **Digital traces** of personal data you leave around (cookies, GPS trackers, the apps you use)
- **(hidden) metadata** in files: sometimes it is not just what you see
- **Where data are stored matters: Schrems vs FISA** (Foreign Intelligence Surveillance Act)
  - Consider the perspective that as an employee of an organization you have accepted that some of your data is transferred overseas, however your study participants did not sign any agreement with third parties
- **Know your rights about your data**
- Familiarize yourself with (and adopt!) **encryption, secure connectivity (MFA, SSH keys, https), secure file sharing and communication (funet, Signal)**

Opening and sharing data must be part of the transparent process of doing research

- **Open science practices** bring **high benefits to researchers** (higher citations, higher impact) and **to the future of science itself** (reproducibility, generalisability, sustainability)
- **“As open as possible, as closed as necessary”**
- **GDPR however can set restrictions on the secondary use of data: opening data for verification or opening data for reuse?**
- **Personal data and long term preservation**

Can anonymised data be opened/shared?

- **Anonymised data are not personal data** so they do not fall under the restriction we have seen before
- However **always consider ethical implications**
  a. What happens when a subject asks to be removed from an open dataset since they can recognize themselves?
  b. What we promise to be anonymous today, might not be anonymous anymore in the future
  c. Although it can be impossible to re-identify an individual, these are still data given by persons who might not be approving re-usage of data for purposes they did not agree with. Anonymous data is often released under CC0; there is no license that forbids large language models to use open data to train the next GPT Language Model
  d. See section "2.6 Why ethics is an important issue in Anonymisation" here:

http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf

Can pseudonymised/minimised data be opened/shared?

- **It always depends on what you have promised to the subjects**
- **Sharing with partner institutions can be done** but it always requires some forms of legal agreements between parties

  Sharing outside EU might not always be possible.
- **Personal data opening:** early days in Europe.
  a. Personal research data should never be fully open, we can make data available on ~~reasonable~~ request as long as the request is NOT handled by the researcher
  b. We need to clarify to the data subject the"secondary use of data". For health and social data in Finland we have the Toisiolaki. EDPB is working on solving issues related to "broad consent".
  c. There are still ethical implications even if the legal side is fine (data leaks, data abuse, data misuse)
  d. Existing and upcoming solution: EGA, SD Submit

Unpopular opinion: should all data be open? should we instead focus more on opening the methods?

- Efforts on opening data for secondary use, only after **data appraisal**: **lawfully and ethically opening personal data is a large coordinated effort between researchers, data stewards, lawyers, ethics, IT infra.**
- **Pre-registration of protocols and registered reports could enhance the research project more than a small open dataset**

Is data protection in EU sending researchers away?

- **Sensitive datasets collected in developing countries (e.g. the largest fingerprint open dataset is from Africa, N = 600)**

Sokoto Coventry Fingerprint Dataset

TL;DR

Take home messages

Take home messages

- You can collect any personal data, but take care of the data protection according to the risks
- Held the highest standard of ethics and law towards your subjects
- Anonymise if possible, minimise otherwise, share on request is fine, but start working on the process before data collection
- (when possible) talk about these issues with the participants, make them comfortable, don’t leave them alone with consent forms and privacy notices

THE END

Thank you!